Loan fraud detection

using

Supervised Learning Classification

Presented by
Gopi Selvaraj

Problem Definition

- All bank branches across India provide MUDRA loans. Such loans have created the low-cost credit concept for micro and small businesses. One of the leading financial institutions in India wants to leverage Machine Learning techniques to determine the client's loan repayment abilities and take proactive steps to reduce the magnitude of exposure to default in future.
- The goal of the problem is to predict whether a client will default on the loan payment or not, given the recent data of all the loan transactions. This can help the institution to distinguish future applicants who might default. For each ID in the Test Dataset, you must predict the "Default" level.

Data dictionary

S.No	Field name	Description	Data type	Data type change
1	Date_Of_Disbursement'	The date on which the lender promises to transfer the sanctioned loan amount to the borrower's account.	Object	datetime64
2	Business	Whether the business is 'Existing' or 'New'	Object	
3	Jobs_Reatained	No. of permanent, full-time employee position that is in existence as of the date of the loan application	Int64	
4	Jobs_Created	No. of permanent, full-time employee position that the business has created as of the date of the loan application	Int64	
5	Year_Of_Commitment	Year in which the loan was sanctioned	Object	Int32
6	Guaranteed_Approved_Loan	Guaranteed approved loan amount.	Object	Float64
7	Borrower_Name	Name of the borrower of the loan	Object	
8	Low_Documentation_Loan	Potential borrower to apply for a mortgage while providing little or no information regarding their employment, income, or assets	Object	
9	Demography	Statistical study of human populations, especially with reference to size and density, distribution, and vital statistics	Object	
10	State_Of_Bank	To which state the bank belongs to.	Object	
11	ChargedOff_Amount	When a borrower defaults on a loan, refers to the amount of money that a lender has declared as unlikely to be collected from the borrower.	Object	Float64
12	Borrower_City	City where the borrower resides	Object	
13	Borrower_State	State where the borrower resides.	Object	
14	Gross_Amount_Balance	Remaining money to be paid to the lender	Object	Float64
15	Count_Employees	No. of employees working in the organization.	Int64	
16	Classification_Code	This may be a unique code that maybe assigned to a loan taken.	Int64	Object
17	Loan_Approved_Gross	Approved loan amount	Object	Float64
18	Gross_Amount_Disbursed	Gross amount paid to the person.	Object	Float64
19	Loan_Term	Time period for the loan in months.	Int64	
20	Commitment_Date	Date within which the loan amount must be paid	Object	datetime64
21	Primary_Loan_Digit	This may be a unique identifier that maybe assigned to a loan taken	Int64	
22	Code_Franchise	Code of the loan provided franchise.	Int64	Object
23	Name_Of_Bank	Indicator whether the card was present during the transaction.	Object	
24	Revolving_Credit_Line	Indicator whether the POS was on the merchant's premises.	Object	
25	Default	Indicator for recurring authorization.	Int64	Object

Shape and distribution of the target variable

- The dataset has **105000** observation & **25** variables.
- The complexity is in finding the solution to the problem based on the chosen techniques, to handle the 'Curse of dimensionality' of the data (considering the data has 25 columns).
- When calculating the distribution of the target variable, we can see the

  Majority class is 'Not-default' / 'False'

  Minority class is 'Default' / 'True'

  with a ratio of **72% : 28%**
- Since the target variable is 'Categorical', we will be building a Supervised Learning Classification model.
- **Python Version:**

  '3.11.7 | packaged by Anaconda, Inc.

Dropping Variables

- None of the variables have been dropped considering the % of the missing values in each of the columns.

Excluded Variables

- Excluding the below variables from the dataset, as they are irrelevant to the target variable and further may introduce unwanted confusion during model building:

DateOfDisbursment	Year_Of_Commitment
Borrower_Name	State_Of_Bank
Borrower_City	Borrower_State
Classification_Code	Code_Franchise
Primary_Loan_Digit	Commitment_Date
Name_Of_Bank	Gross_Amount_Balance

Missing Value Imputation

- There were 5 currency columns where 'Rs.' which was removed from the column values & data type 'Object' converted to 'Float'
  - Guaranteed_Approved _Loan
  - ChargedOff_Amount
  - Gross_Amount_Balance
  - Loan_Approved_Gross
  - Gross_Amount_Disbursed

- There was 1 column which was renamed appropriately.
  - Jobs_Reatained : 'Jobs_Retained'

- There were 2 columns with data type 'Object' converted to 'Datetime'
  - Date_Of_Disbursement
  - Commitment_Date

- There were 3 columns with data type 'Int' converted to 'Object' (categorical variable)
  - Classification_Code
  - Code_Franchise
  - Default (Target variable)

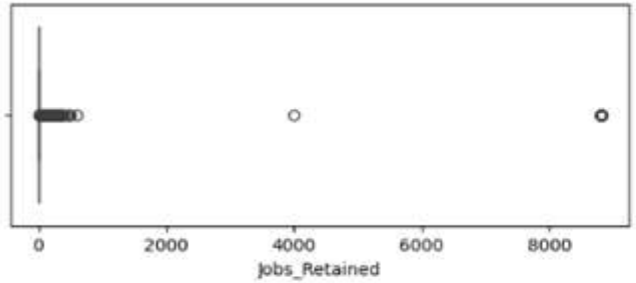
- There were 2 columns with mismatched values which have been rectified
  - 'Year_Of_Commitment' : 1976A -> 1976
  - 'Borrower_State' : Removed white spaces within Subclasses

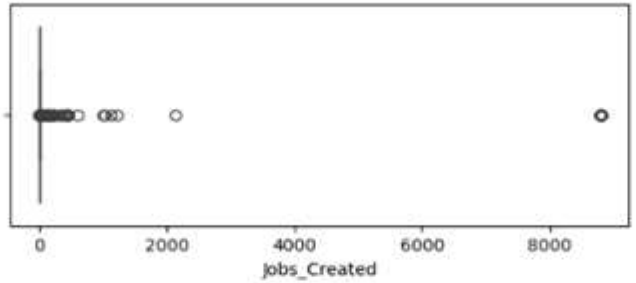
- There was 1 column with data type 'Object' converted to 'Int'
  - 'Year_Of_Commitment'

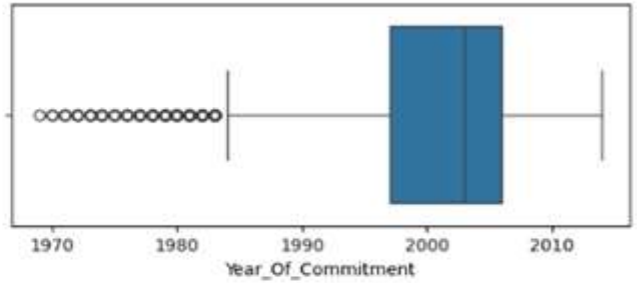
Exploratory Data Analysis

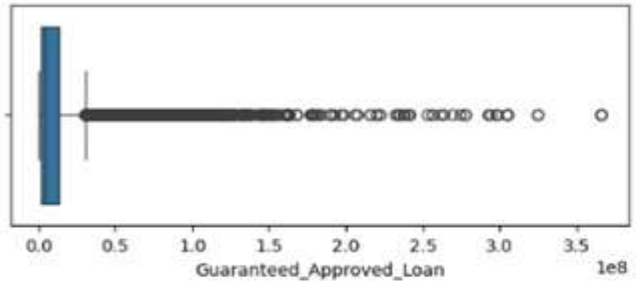
Univariate analysis - Numeric variables

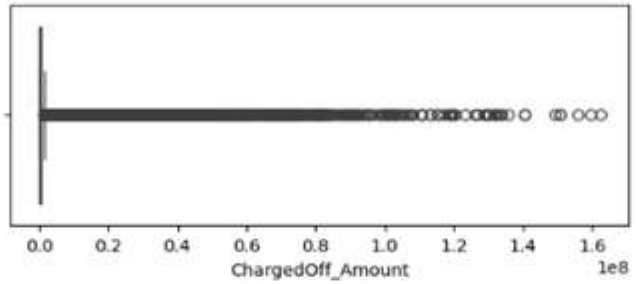
BOX PLOT: (Detection of IQR, Minimum, Maximum, 25%, 50%, 75% & outliers in a numeric column)

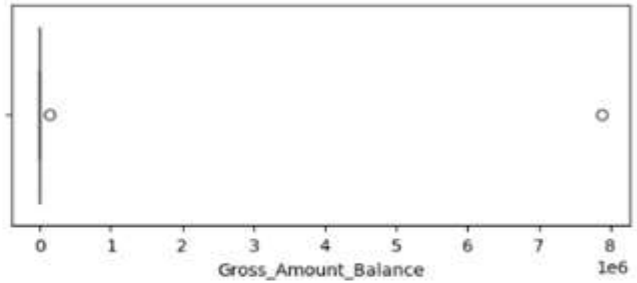


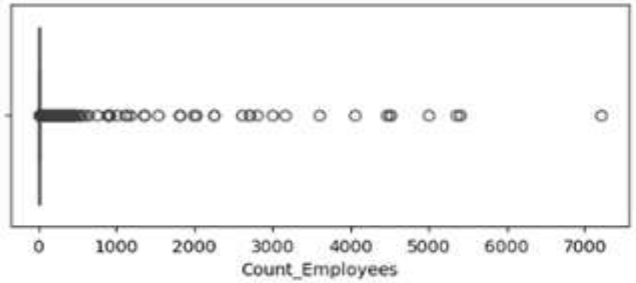


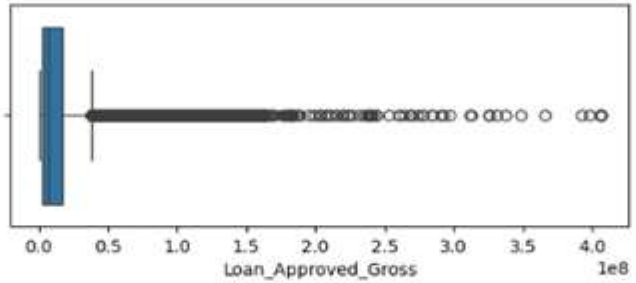


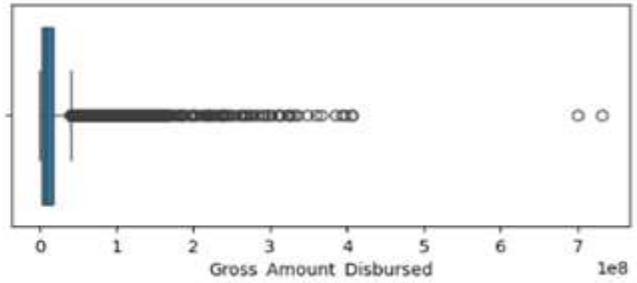


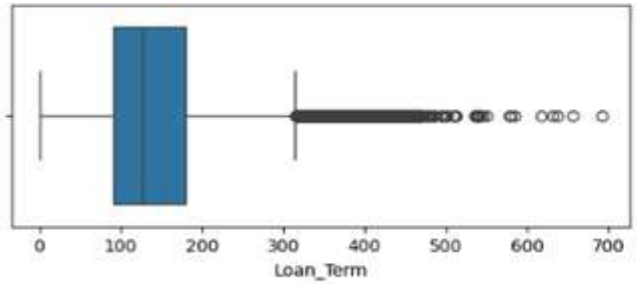


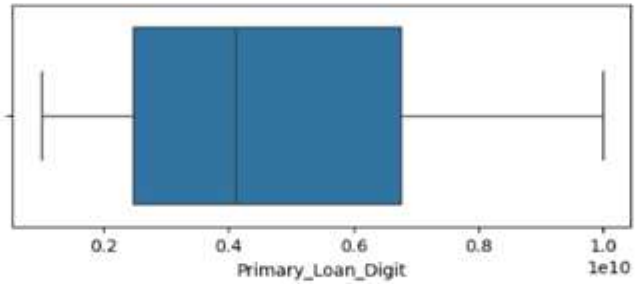




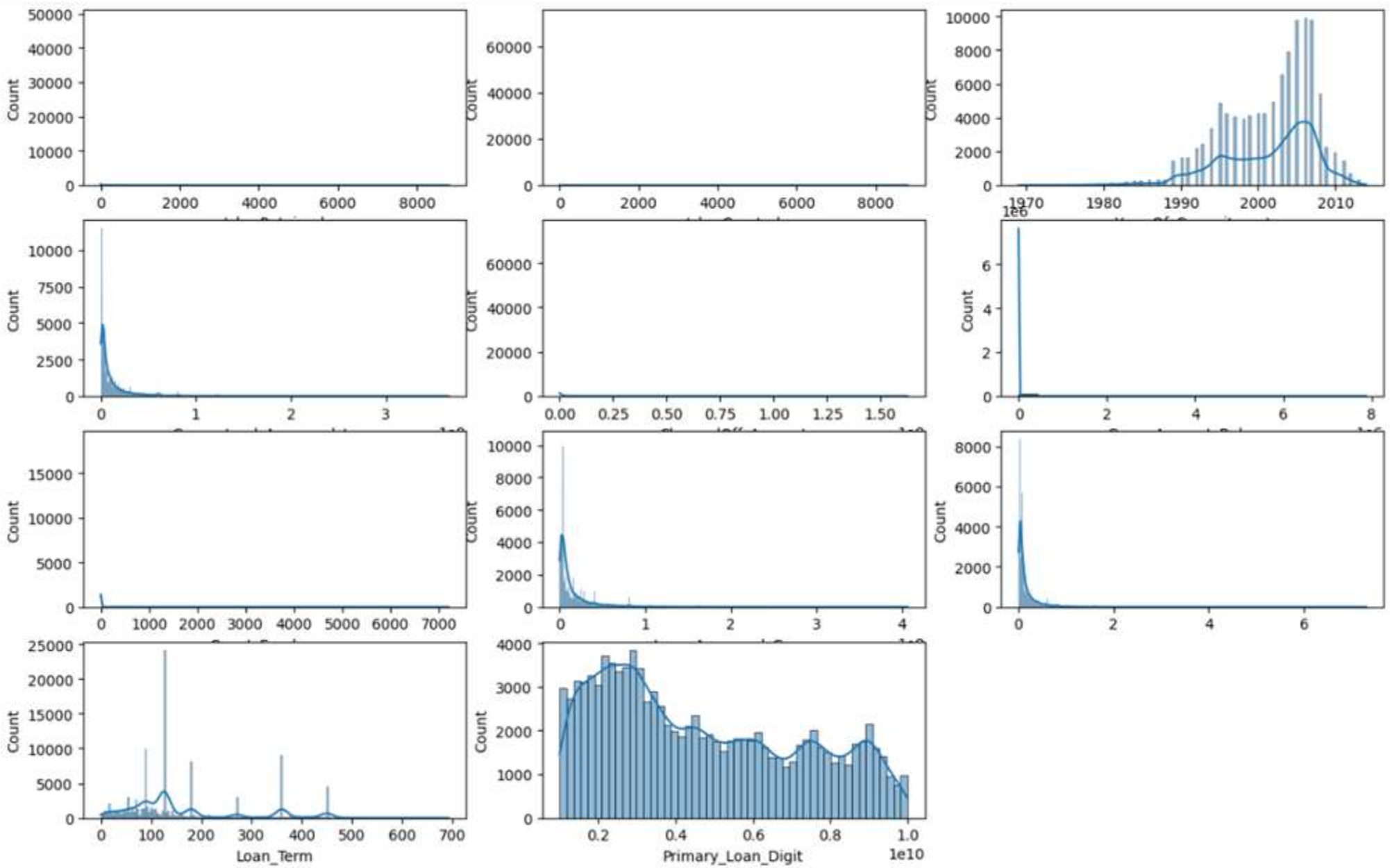








HISTOGRAM: (Detect skewness; visualize the distribution of the data; detect the position of mean median & mode)



Inference:

'Jobs_Retained'

- Distribution of data : Extreme positive skewness (32.33), indicating presence of outliers the in higher scale.
- 50% of the values lie around 4
- We can observe huge no. of outliers & extreme outlier values of 4000 & 8800

'Jobs_Created'

- Distribution of data : Extreme positive skewness (32.41), indicating presence of outliers in the higher scale.
- 50% of the values lie around 1
- We can observe huge no. of outliers & extreme outlier values of 2100 & 8800

'Year_Of_Commitment'

- Distribution of data : Moderate negative skewness (-0.75), indicating presence of outliers in the lower scale.
- 50% of the values lie around 2004
- We can observe huge no. of outliers & an extreme outlier value of 1969

'Guaranteed_Approved_Loan'

- Distribution of data : Extreme positive skewness (3.8), indicating presence of outliers in the higher scale.
- 50% of the values lie around 11541760
- We can observe huge no. of outliers & an extreme outlier value of 365760000

'ChargedOff_Amount'

- Distribution of data : Extreme positive skewness (8.9), indicating presence of outliers in the higher scale.
- 50% of the values lie around 617016
- We can observe huge no. of outliers & an extreme outlier value of 162559918.7

'Gross_Amount_Balance’

- Distribution of data : Extremely positively skewed (323.8), indicating presence of outliers in the higher scale.
- Almost all of the values are 0 and only two other values are present (143052.8 & 7876682.2)
- We can observe an extreme value of 7876682.2

'Count_Employees’

- Distribution of data : Extremely positively skewed (68), indicating presence of outliers in the higher scale.
- 50% of the values lie around 7
- We can observe an extreme value of 7200

'Loan_Approved_Gross’

- Distribution of data : Extremely positively skewed (3.7), indicating presence of outliers in the higher scale.
- 50% of the values lie around 14305280
- We can observe huge no. of outliers & an extreme outlier value of 162559918.7

‘Gross_Amount_Disbursed’

- Distribution of data : Extremely positively skewed (4.2), indicating presence of outliers in the higher scale.
- 50% of the values lie around 14874585.4
- We can observe huge no. of outliers & an extreme outlier value of 731113600

‘Loan_Term’

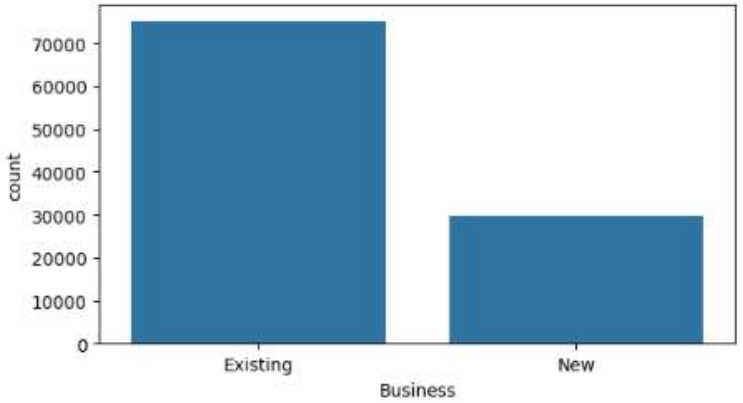
- Distribution of data : Extremely positively skewed (1.2), indicating presence of outliers in the higher scale.
- 50% of the values lie around 90
- We can observe huge no. of outliers & an extreme outlier value of 692

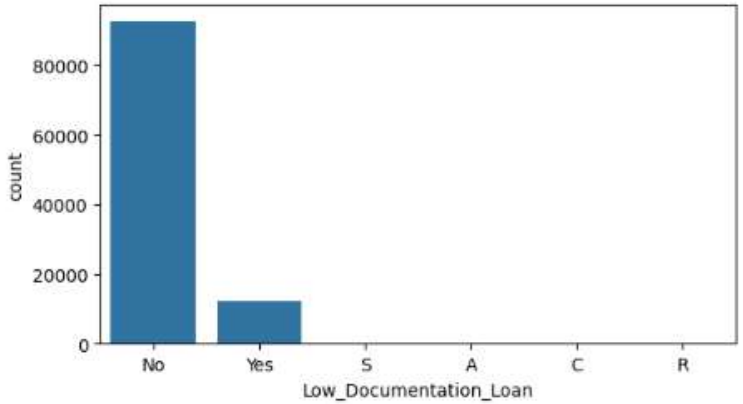
'Primary_Loan_Digit’

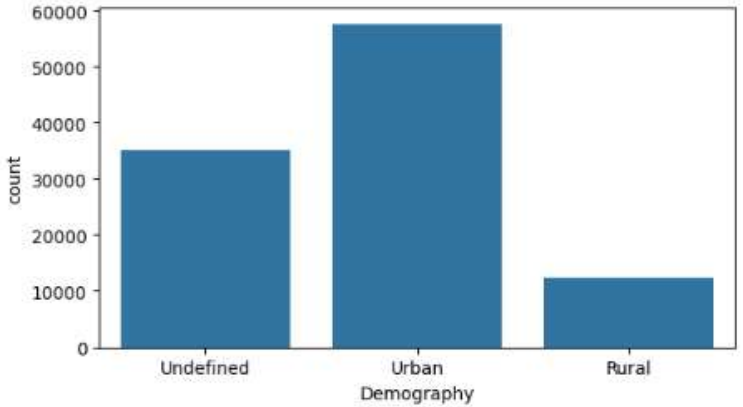
- Distribution of data : Near normal distribution & no outliers (no skewness)
- 50% of the values lie around 4282320745.2- We can't observe any outliers.
- This column doesn't make sense as a predictor variable because it is just a unique identifier assigned to each loan taken.

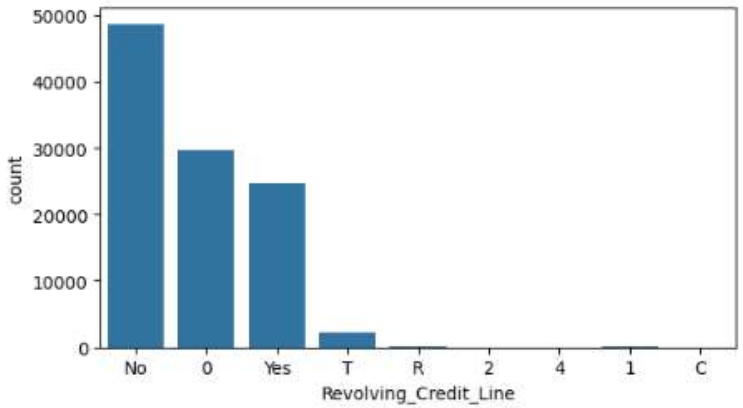
Univariate analysis - Categorical variables

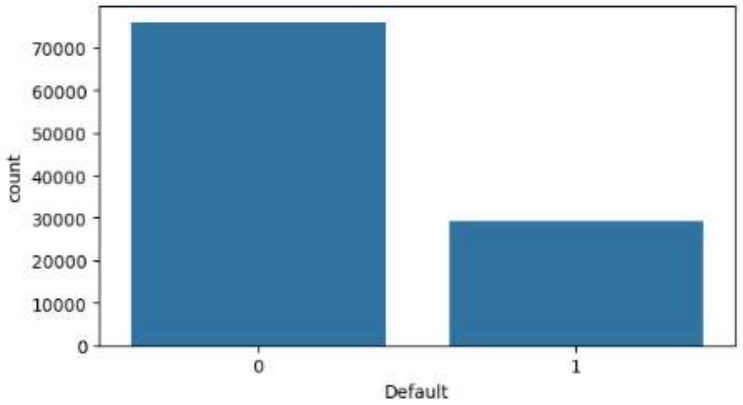
COUNT PLOT: (Visualize each of the subclasses total count in the form of a bar chart)





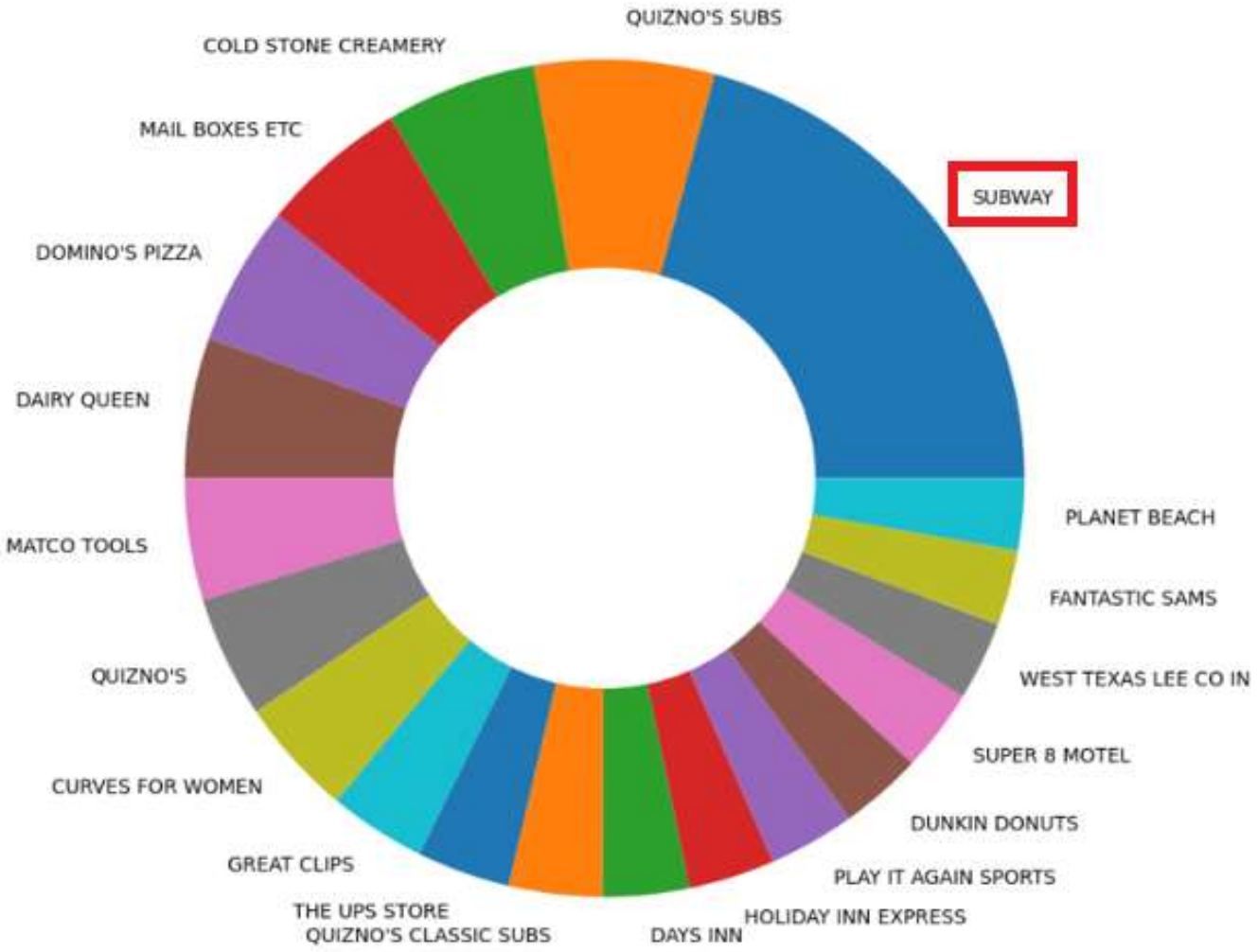




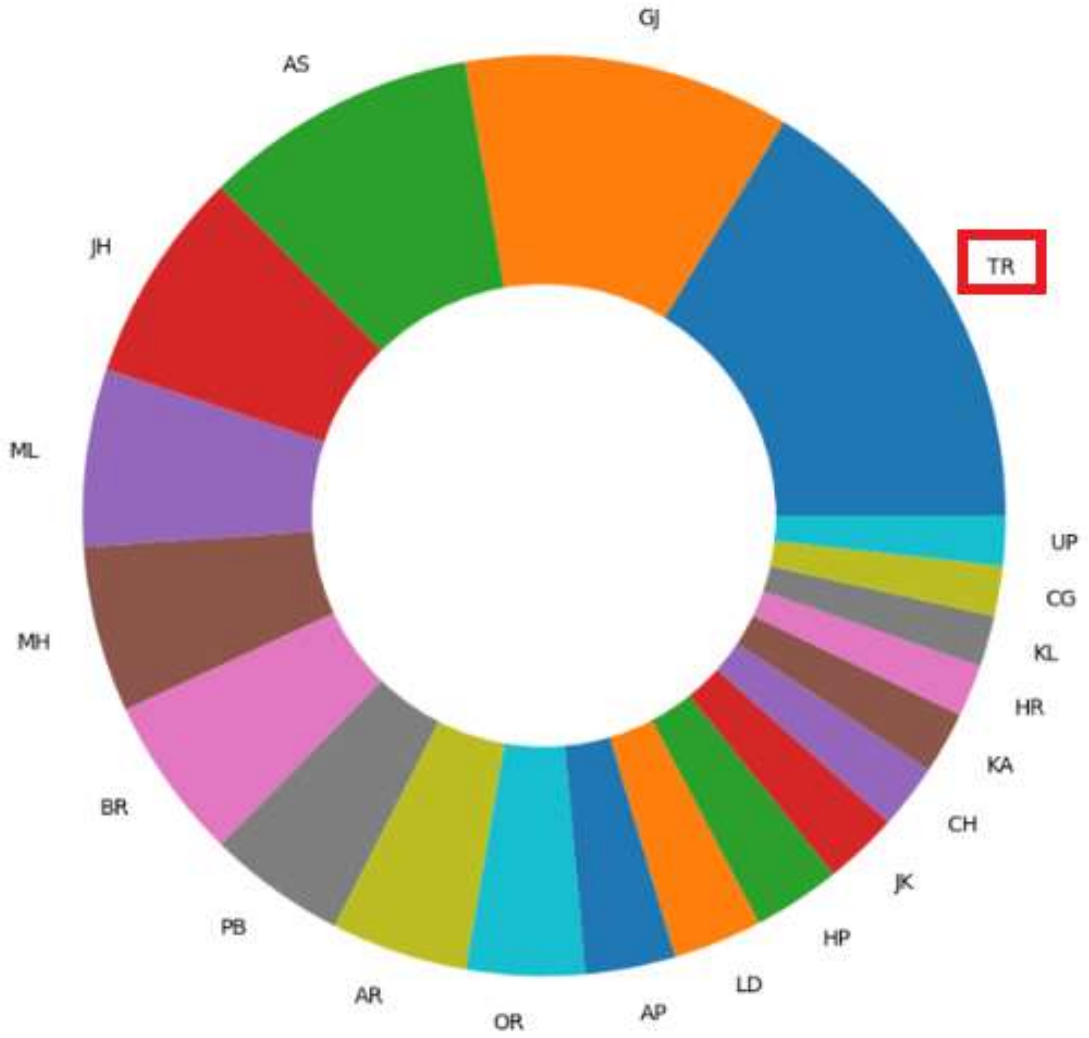


PIE PLOT: (Considering the huge no. of subclasses for the below variables we have decided to use a Pie plot to represent the top 10 sub-classes and visualized them)

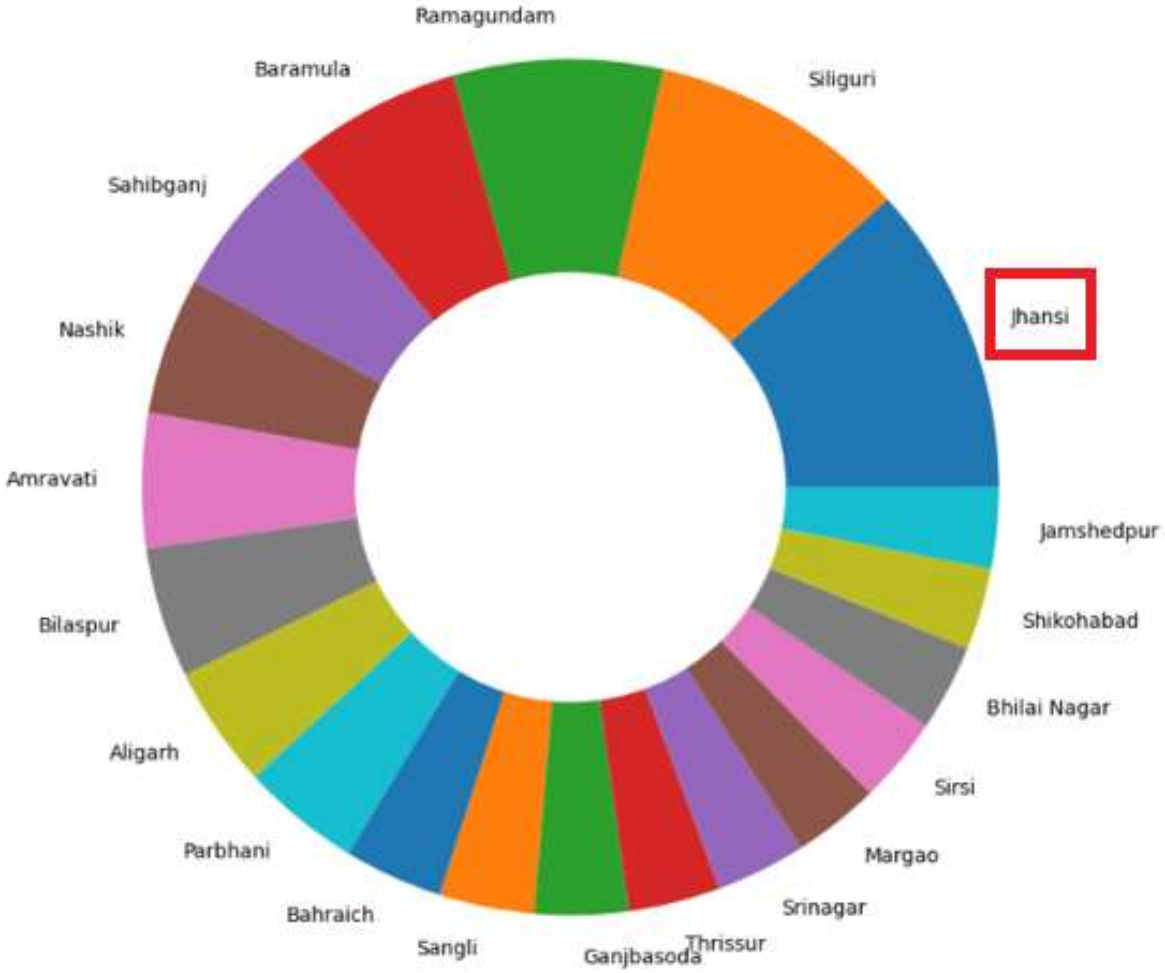
‘Borrower_Name’



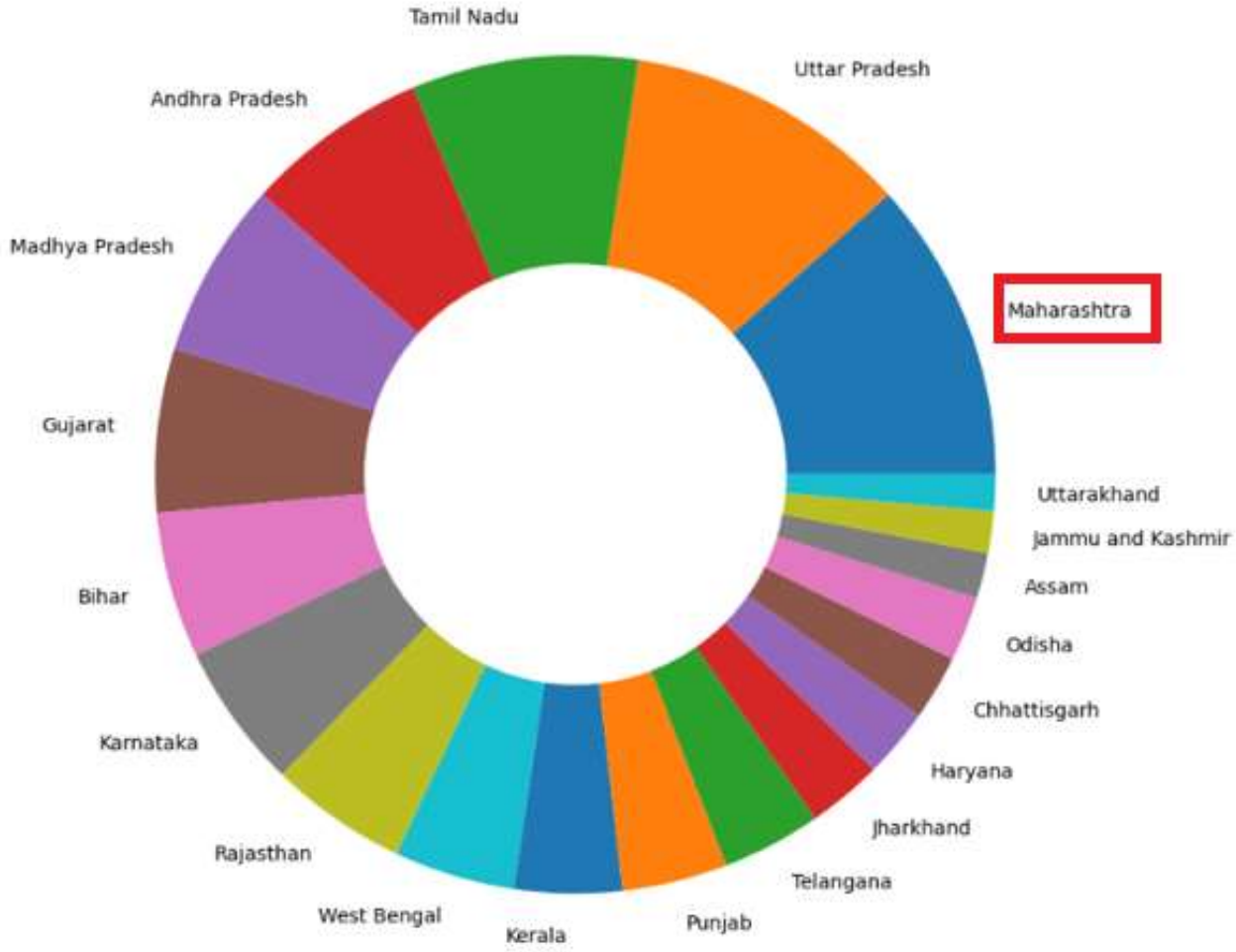
‘State_Of_Bank’



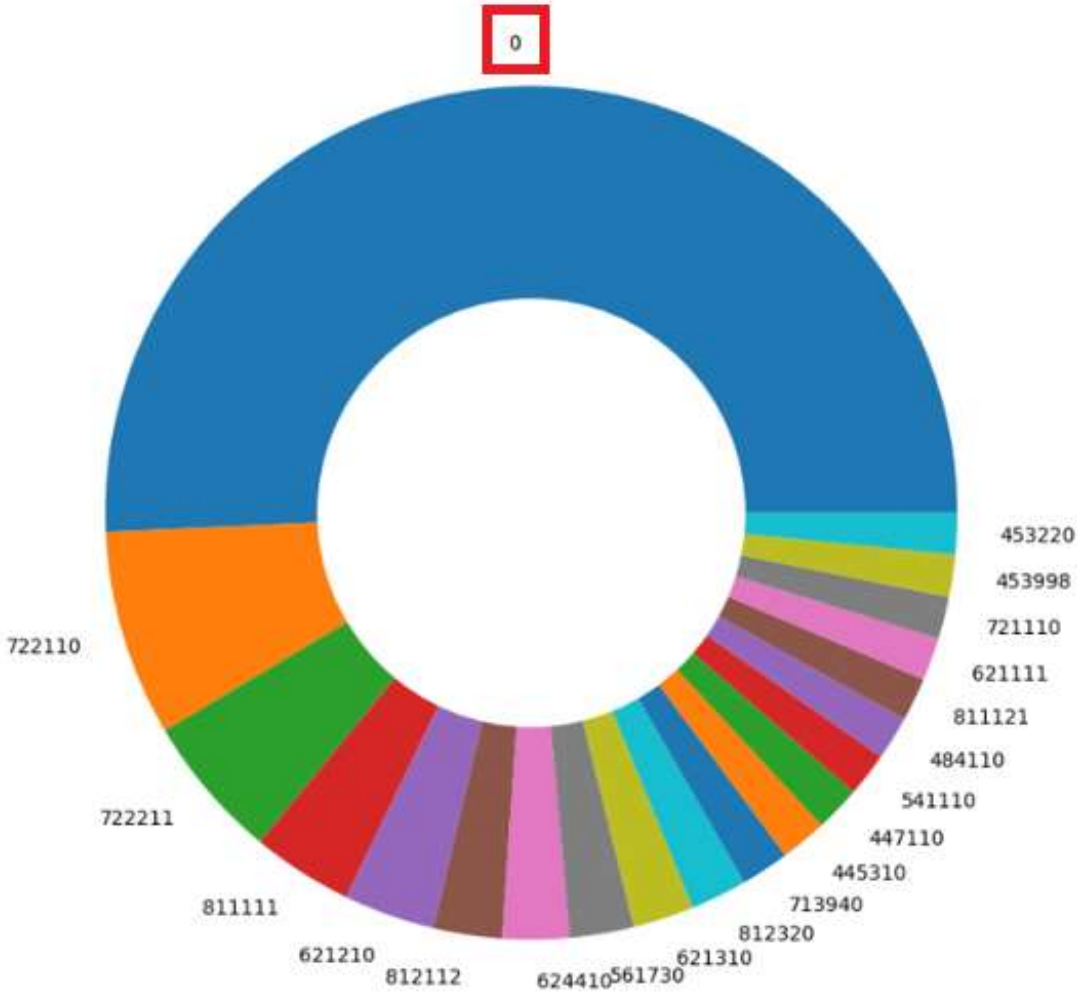
‘Borrower_City’



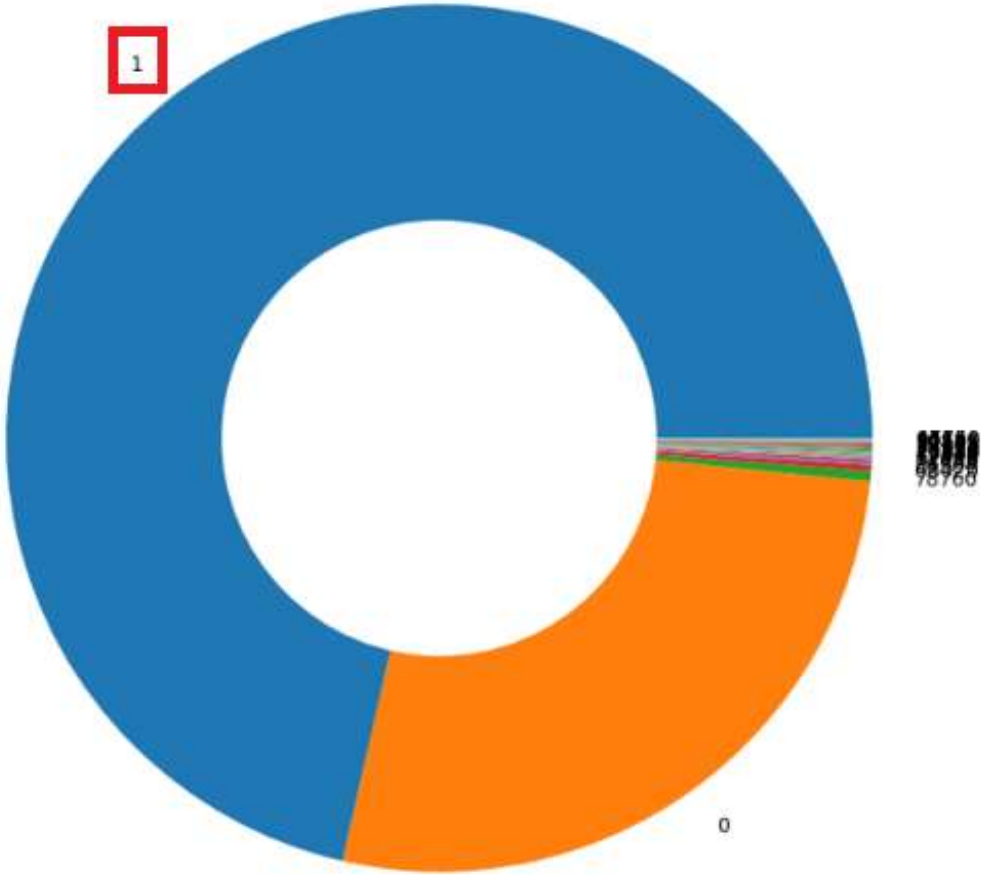
‘Borrower_State’



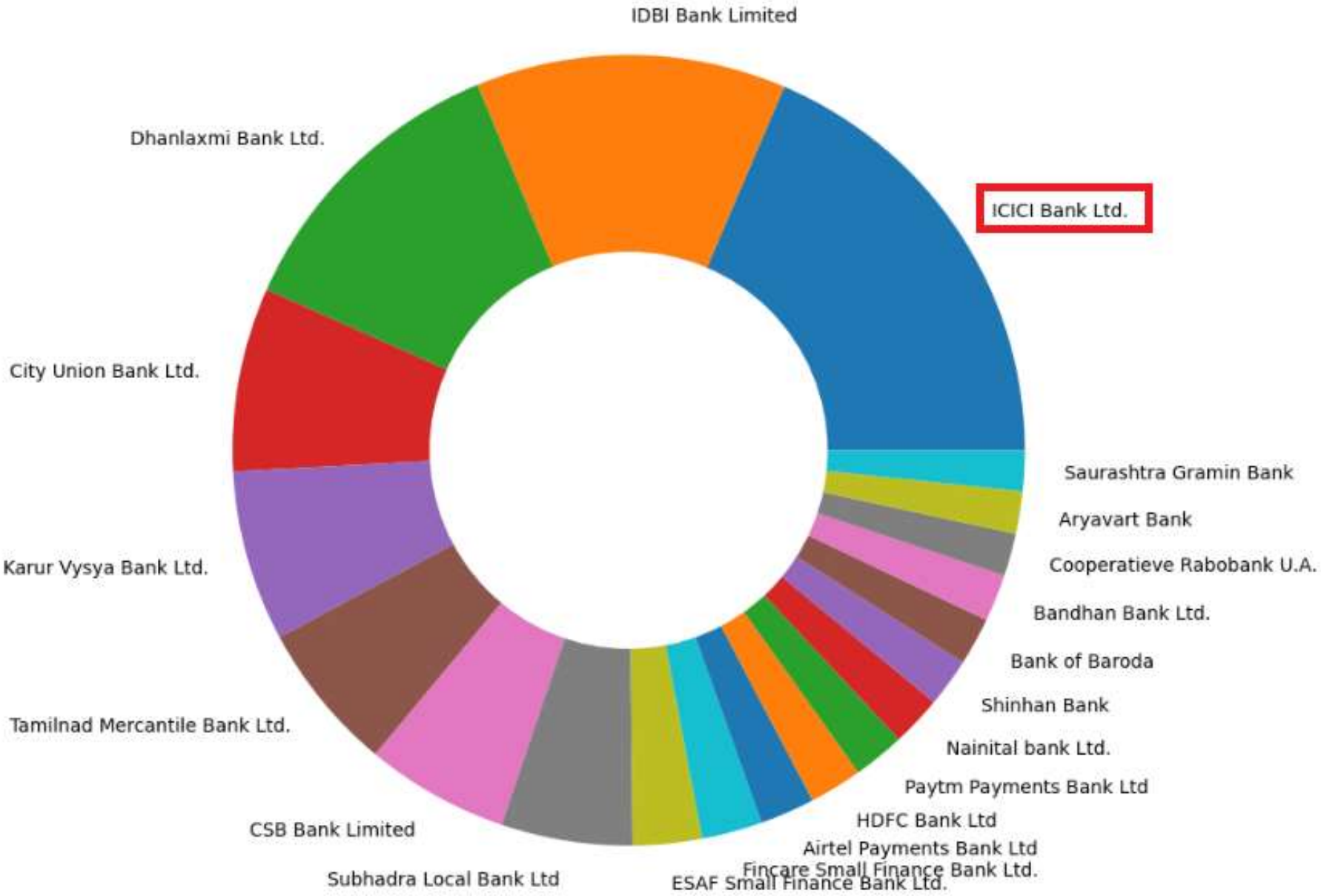
‘Classification_Code’



‘Code_Franchise’



‘Name_Of_Bank’


IDBI Bank Limited
Dhanlaxmi Bank Ltd.
ICICI Bank Ltd.
City Union Bank Ltd.
Saurashtra Gramin Bank
Aryavart Bank
Cooperatieve Rabobank U.A.
Karur Vysya Bank Ltd.
Bandhan Bank Ltd.
Bank of Baroda
Shinhan Bank
Tamilnad Mercantile Bank Ltd.
Nainital bank Ltd.
Paytm Payments Bank Ltd
HDFC Bank Ltd
CSB Bank Limited
Airtel Payments Bank Ltd
Fincare Small Finance Bank Ltd.
Subhadra Local Bank Ltd
ESAF Small Finance Bank Ltd.

Inference:

'Business'
- Most of the loans were taken by businesses which are already 'Existing' (75211)
- Least no. loans were taken by business which are 'New' (29789)
- There is high imbalance in subclasses & bias towards 'Existing'

'Low_Documentation_Loan'
- Most of the loans had opted 'No' (92675) for 'Low_Documentation_Loan'
- Least no. of values are 'R' (6)
- There is high imbalance in subclasses & bias towards 'No'

'Demography'
- Most of the loans were taken from 'Urban' areas (57598)
- Least no. of loans were taken from 'Rural' areas (10000)
- There is high imbalance in subclasses & is bias towards 'Rural'

'State_Of_Bank'
- Most of loans were provided by banks from the state 'TR' (15552)
- Least no. of loans were provided by banks from the state 'MN' (18)

'Borrower_City'
- Most of the loans are taken from the city 'Jhansi' (1504)
- Least no. of loans are taken from the city 'Mudhol' (19)

'Borrower_State’
- Most of the loans were taken by people belonging to the states of 'Maharashtra' (11699) & 'Uttar Pradesh' (11101)
- Least no. of loans were taken by people belonging to the state 'Bulandshahr' (34)

'Classification_Code’
- Most of the loans belonged to classification code '0' (22209)
- This column doesn't make sense in terms of actual predictor variable, but rather to be an unique identifier.

'Code_Franchise’
- Most of the loans belonged to code franchise '1' (71874)
- This column doesn't make sense in terms of actual predictor variable, but rather to be an unique identifier.

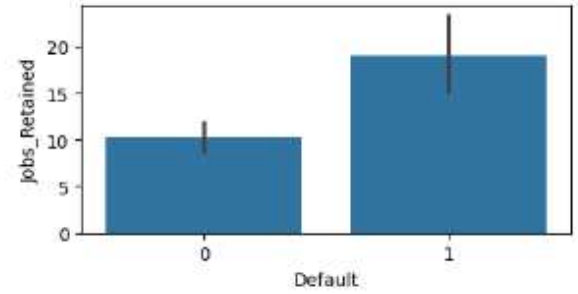
'Name_Of_Bank’
- Most of the loans were provided by 'ICICI Bank Ltd.' (11215)
- Least no. of loans were provided by 'Barclays Bank Plc.' (93)

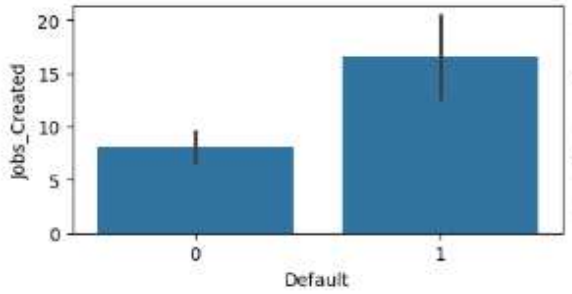
'Revolving_Credit_Line’
- Most of the values belong to 'No' (48616)
- Least no. of values belong to '2' , '4', 'C' (1)

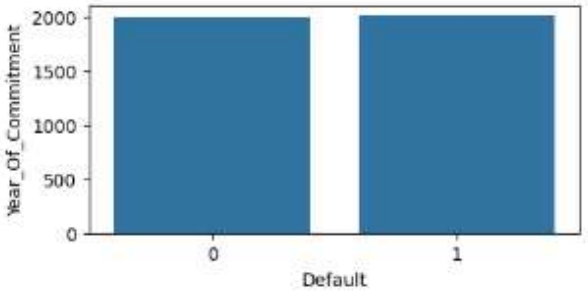
'Default’
- Most of the loans were ‘Not-default’ ['0’ = 75896] (72%)
- Least no. the loans were ‘Default’ ['1’ = 29104] (28%)

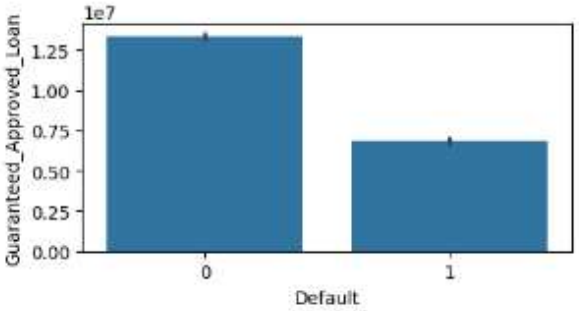
Bivariate analysis – Numerical VS Categorical

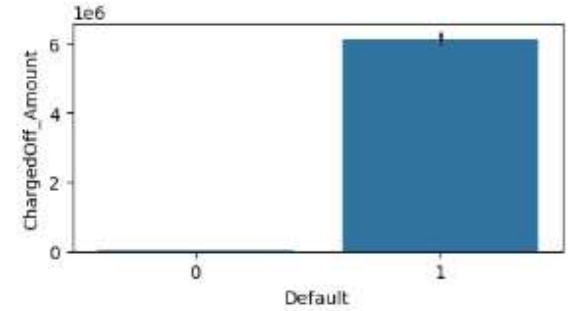
BAR PLOT: [Seaborn bar plot gives the mean of each subclass (of the target variable) in regards to the numeric variable]

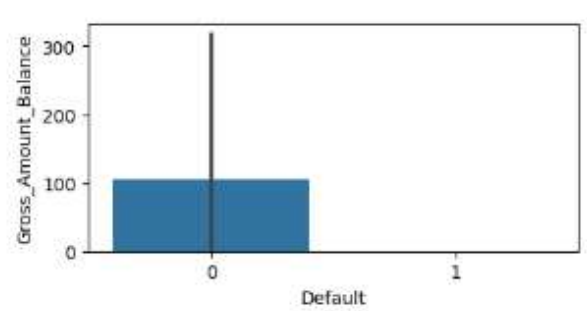


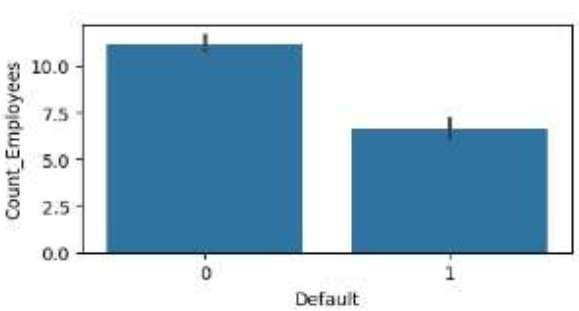


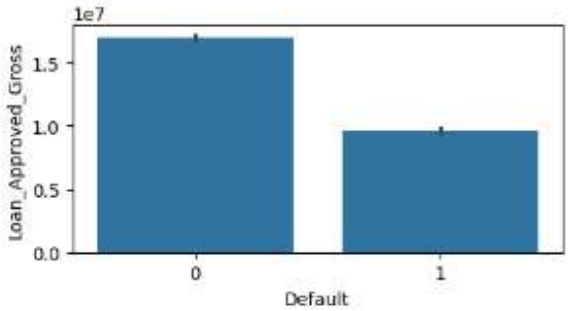


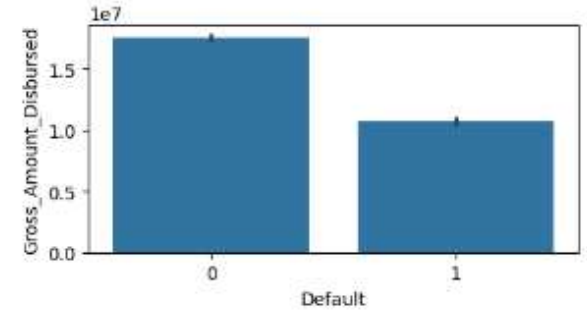


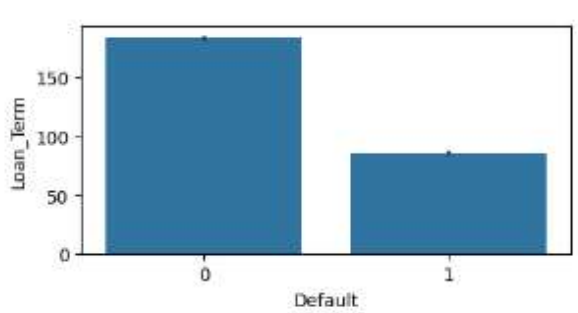


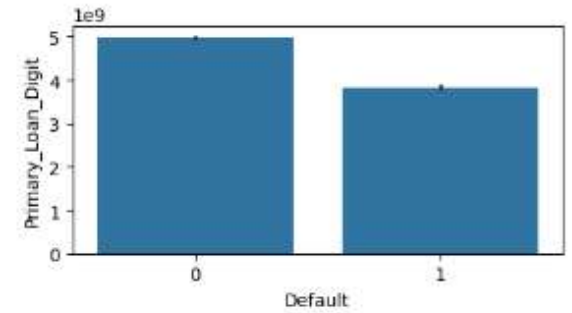




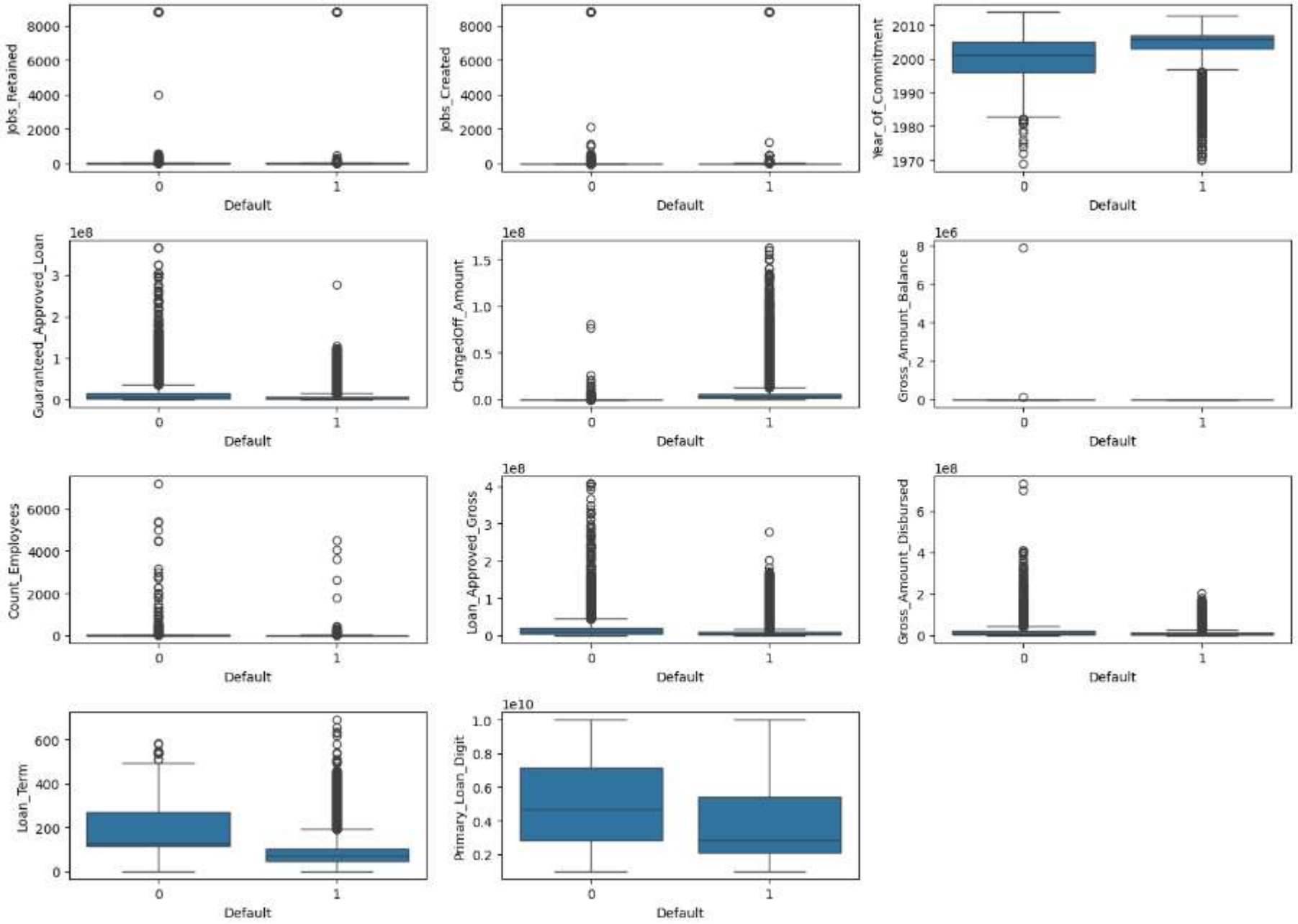








BOX PLOT: (Detection of IQR, distribution, Minimum, Maximum, 25%, 50%, 75%, outliers of each subclass of the target variable in regards to the numeric variable)



Inference:

'Jobs_Retained' VS 'Default'

- On average we can see companies who have retained jobs have more 'Default' (18.9) than companies with 'Not-default' (10.2)
- This may be due to the company needed to payout the salary of the employees but didn't return a profit on their respective business.
- 50% of the data 'Not-default' lies around 4 & 'Default' lies around 5.
- Similar distribution of data between the target variable subclasses.
- There seems to be relationship between these two variables.

'Jobs_Created' VS 'Default'

- On average we can see companies who have created jobs have more 'Default' (16.5) than companies with 'Not-default' (8)
- This may be due to the company which has created more jobs & is required payout the salary of the employees but didn't return a profit on their respective business.
- 50% of the data 'Not-default' lies around 1 & 'Default' lies around 2.
- Similar distribution of data between the target variable subclasses
- There seems to be relationship between these two variables.

'Year_Of_Commitment' VS 'Default'

- Considering the average, there doesn't seem to be significant relationship between 'Year_Of_Commitment' & 'Default'
- 50% of the data 'Not-default' lies around 9 & 'Default' lies around 4

'Guaranteed_Approved_Loan' VS 'Default'

- On average we can see people who have got high 'Guaranteed_Approved_Loan' have 'Not-default' (6822169) but people with lower value have 'Defaults' (13317860)
- 50% of the data 'Not-default' lies around 13817600 & 'Loan default' lies around 5486400
- There seems to be relationship between these two variables.

'ChargedOff_Amount' VS 'Default'

- On average we can see people who have got high 'ChargedOff_Amount' have 'Defaults' (6146350) & low 'ChargedOff_Amount' have 'Not-default' (11430.1)
- 50% of the data 'Not-default' lies around 0 & 'Default' lies around 4797592.6
- There seems to be relationship between these two variables.

'Gross_Amount_Balance' VS 'Default'

- On average we can see people with high 'Gross_Amount_Balance' have 'Not-default' (105.6) & people with low 'Gross_Amount_Balance' have 'Default' (6.5)
- There seems to be relationship between these two variables.

'Count_Employees' VS 'Default'

- On average we can see companies with more employees have 'Not-default' (11.1) while companies with low no. of employees have 'Default' (6.5)
- 50% of the data 'Not-default' lies around 8 & 'Default' lies around 4
- There seems to be relationship between these two variables.

'Loan_Approved_Gross' VS 'Default’

- On average we can see companies who got high 'Loan_Approved_Gross' have ‘Not-default' (16954400) compared to people with low 'Loan_Approved_Gross' who have ‘Default’ (9622007).
- 50% of the data ‘Not-default' lies around 17068800 & ‘Default' lies around 6096000
- There seems to be relationship between these two variables.

'Gross_Amount_Disbursed' VS 'Default’

- On average people with high 'Gross_Amount_Disbursed' have ‘Not-default' (17536700) while people with 'Gross_Amount_Disbursed' have 'Default' (10689470)
- 50% of the data 'Not loan default' lies around 16662400 & ‘Default' lies around 9207093.6
- There seems to be relationship between these two variables.

'Loan_Term' VS 'Default’

- On average people with high 'Loan_Term' have ‘Not-default’ (183.5) while people with low 'Loan_Term' have 'Default' (85.6)
- 50% of the data ‘Not-default' lies around 153 & ‘Default' lies around 59.

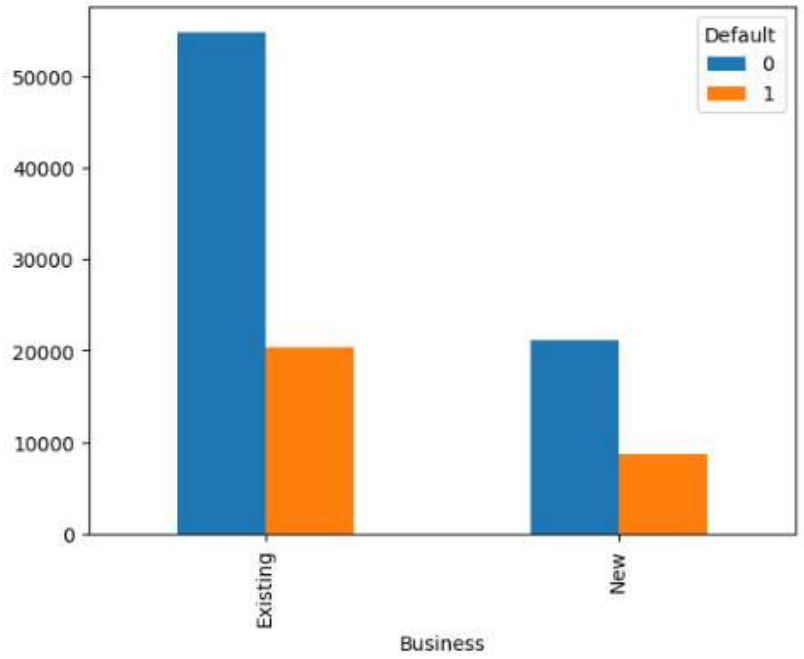
'Primary_Loan_Digit' VS 'Default’

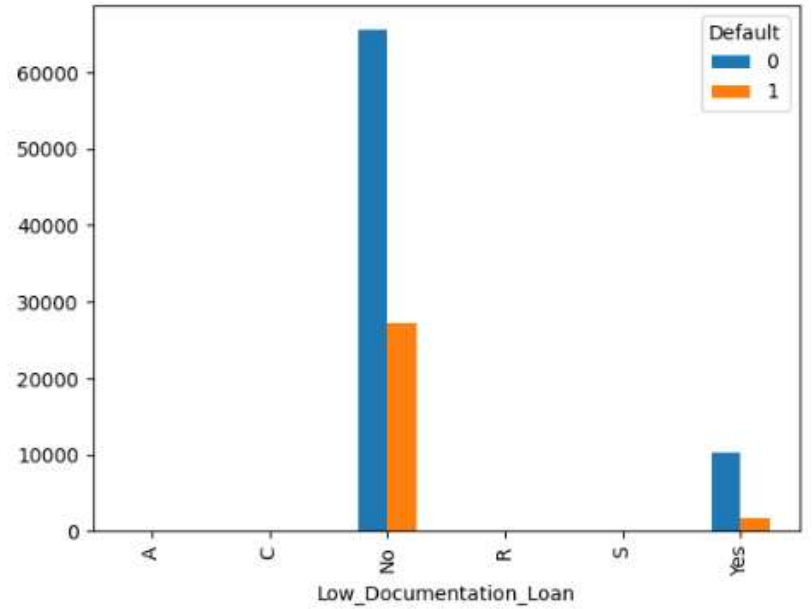
- This column can be just a unique identifier that we may consider discarding that might not be a good predictor variable.

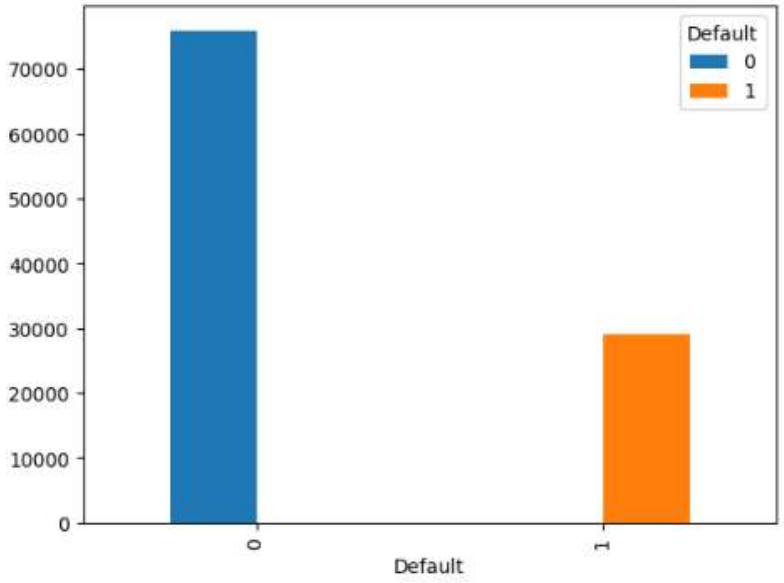
Bivariate analysis

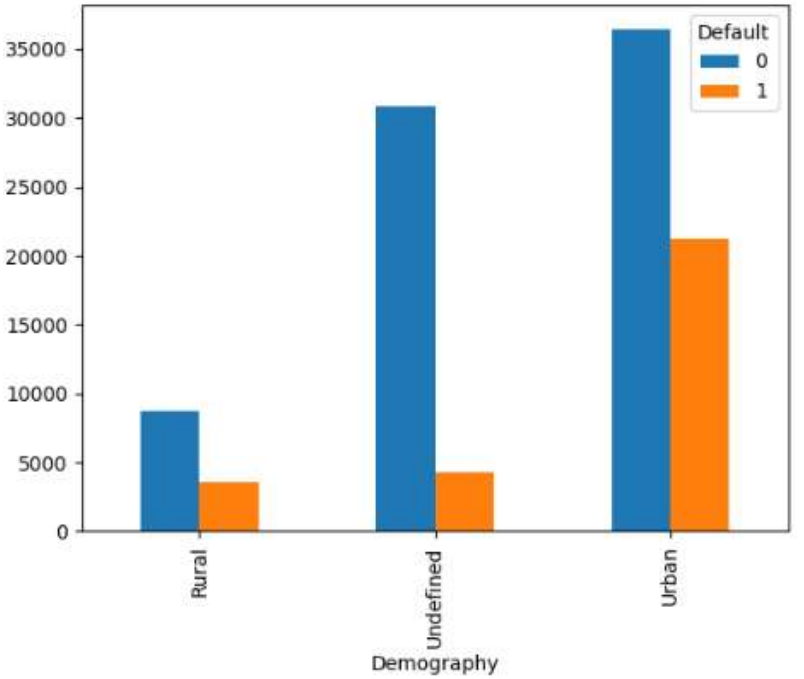
Categorical VS Categorical

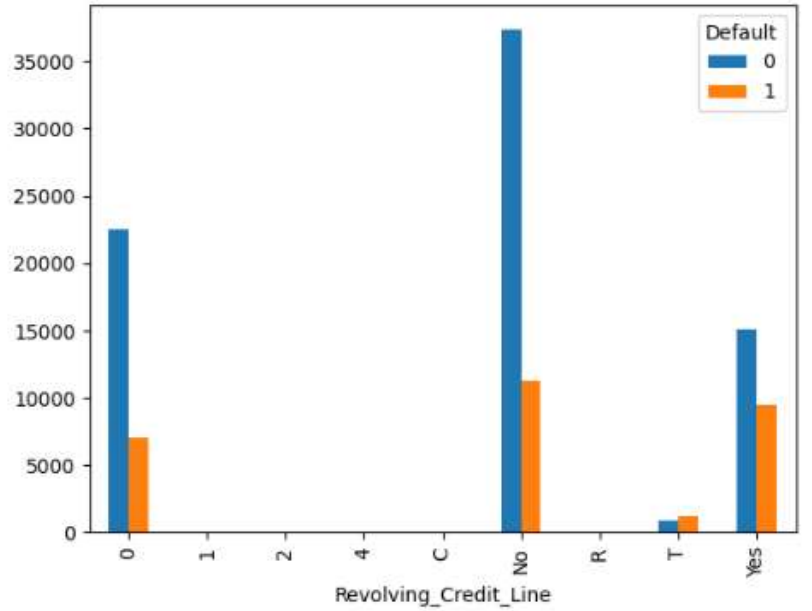
CROSS TABULATION
(Understand the relationship between two categorical variables)











Inference:

'Business' VS 'Default':

- We can see most of the existing 'Business' (20386) have 'Default' compared to the 'New' business who have less 'Default' (8718)
- There seems to be a relationship between the two variables.

'Low_Documentation_Loan' VS 'Default':

- We can see most of the companies who "Don't" have 'Low_Documentation' have 'Default' (27191).
- We can also see considerable no. of companies 'Have' 'Low_Documentation' have 'Default' (1803).
- There seems to be a relationship between the two variables.

'Demography' VS 'Default':

- We can see most of the 'Urban' (21223) companies have 'Default' compared to the 'Rural' (3613) companies.

'Revolving_Credit_Line' VS 'Default' :

- We can see most of the companies who dont have a 'Revolving_Credit_Line' have the highest no. of 'Default' (11312) followed by companies who 'Have a' 'Revolving_Credit_Line'(9495).
- Third place is held by '0' subclass (7055)

Multivariate analysis

Correlation matrix : (Visualize the correlation between the numeric variables)

Inference:

We can infer the presence of Multi-collinearity (existence of correlation between the independent / 'x' variables) from the correlation matrix.

We can observe high positive correlation between:

- 'Loan_Approved_Gross' & 'Guaranteed_Approved_Loan' (0.98)
- 'Gross_Amount_Disbursed' & 'Guaranteed_Approved_Loan' (0.94)

We can observe moderate positive correlation between:

- 'Loan_Term' & 'Guaranteed_Approved_Loan' (0.53)
- 'Loan_Term' & 'Loan_Approved_Gross' (0.5)
  'Loan_Term' & 'Gross_Amount_Disbursed' (0.47)

We can observe weak positive correlation between:

- 'Loan_Approved_Gross' & 'Charged_off_Amount' (0.26)
- 'Gross_Amount_Disbursed' & 'Charged_off_Amount' (0.26)

We can observe weak negative correlation between:

- 'Loan_Term' & 'Year of commitment' (-0.31)
- 'Primary_Loan_Digit' & 'Year of commitment' (-0.21)

Statistical test of significance:

Since the data has failed to meet the assumptions

(i) Data normality (Shapiro test)

(ii) Data has equal variance (Levene test)

we have opted for Non-parametric test to try to understand if any relationship exists between the independent variables and dependent variable.

Numeric VS Categorical – Mann -Whiteney U (Non-Parametric)

Categorical VS Categorical – Chi square contingency test (Non-Parametric)

Statistically significant numeric columns:

- 'Jobs_Retained'
- 'Jobs_Created'
- 'Guaranteed_Approved_Loan'
- 'ChargedOff_Amount'
- 'Count_Employees'
- 'Loan_Approved_Gross'
- 'Gross_Amount_Disbursed'
- 'Loan_Term'

Statistically significant categorical columns:

- 'Business'
- 'Low_Documentation_Loan'
- 'Demography'
- 'Revolving credit line'

Data preprocessing

Scaling the data

- Scaling is performed in the dataset for the numeric variables after train test split as we can hide the mean standard deviation of training data from the test data.
- Scaling has been performed as part of one of Yeo-Johnson transformation (which performs both scaling & transformation)

Outlier treatment:

- The popular methods of handling the outliers are Trimming/removing the outlier based on IQR or z-Score or capping them.
- But, we would like to consider the outliers in this dataset as a pattern and prefer not to treat them but to handle them differently since it is important for the model to get trained based on some extreme values in order to predict in an efficient way consider their nature in the financial sector.

Transformation technique:

- We prefer transformation of the numeric variables (using Yeo-Johnson transformation) so that we can try to convert them to near normal distribution.

Encoding the Categorical Variables:

- The performance of a machine learning model not only depends on the model and the hyperparameters but also on how we process and feed different types of variables to the model. Since most machine learning models only accept numerical variables, pre-processing the categorical variables becomes a necessary step. Converting these categorical variables to numbers such that the model is able to understand and extract valuable information is known as Encoding. There are various encoding techniques available like Dummies, One Hot Encoder, Label Encoder, Ordinal Encoder etc., We have used Label Encoder for encoding our categorical variables considering the no. of categorical variables and we have also seen there is not a significant improvement in model performance when dummy encoding was utilized.

- We have utilized **Label encoding technique**. In this case, retaining the order is important. Hence encoding should reflect the sequence. In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person likewise.

- The final processed data which we would be using in building various Classification Models like Logistic Regression, Decision Tree, Random Forest etc. We, with the help of the built models we would infer on the significance and effects of each independent variable on our target variable for predicting the patterns and rate of successful conversion to give some insightful ideas for effective marketing.

Model Building & Evaluation

- The Modeling is the core of any machine learning project. This step is responsible for the results that should satisfy or help satisfy the project goal.

- Building a model in machine learning is creating a mathematical representation by generalizing and learning from training data.

- Our problem statement come under the Classification thus we have decided to use various models namely

1. Logistic Regression Model
2. Decision Tree Model
3. Random Forest Model
4. Ada Boost Technique
5. Gradient Boosting Technique
6. Extreme Gradient Boosting Technique

- **Evaluation metrics**: Recall, Precision, Accuracy & F1 Score (weighted). Due to imbalance in the subclass of the target variable we have opted for weighted evaluation metrics which will give us the correct representation of the metrics thereby will estimate the model performance appropriately.

Hyper tuning the model

Model: Decision Tree Classifier

GridSearchCV (run for below parameters)

Criterion	Entropy, Gini
Max Depth	30, 35, 40
Min Samples Split	50, 60, 70, 80, 90
Min Samples Leaf	10, 20, 30, 40

Best Parameters

Criterion	Entropy
Max Depth	30
Min Samples Split	80
Min Samples Leaf	10

Model performance – Test & Training scores

	Model	Accuracy	Recall	Precision	F1 Score	TN	FN	FP	TP
0	Logistic Regression [Base model] - Test	0.984190	0.984190	0.984182	0.984130	15084	237	95	5584
1	Logistic Regression [Base model] - Train	0.984393	0.984393	0.984381	0.984336	60330	924	387	22359
2	Logistic Regression [VIF columns] - Test	0.980190	0.980190	0.980299	0.980045	15104	341	75	5480
3	Logistic Regression [VIF columns] - Train	0.979881	0.979881	0.979973	0.979737	60391	1364	326	21919
4	Logistic Regression [Significant columns] - Test	0.984190	0.984190	0.984182	0.984130	15084	237	95	5584
5	Logistic Regression [Significant columns] - Train	0.984393	0.984393	0.984381	0.984336	60330	924	387	22359
6	Logistic Regression [Transformed data] - Test	0.993905	0.993905	0.993984	0.993920	15068	17	111	5804
7	Logistic Regression [Transformed data] - Train	0.994036	0.994036	0.994087	0.994047	60323	107	394	23176
8	Decision Tree [normal data & hyper-parameter t...	0.993905	0.993905	0.993984	0.993920	15068	17	111	5804
9	Decision Tree [normal data & hyper-parameter t...	0.994048	0.994048	0.994098	0.994059	60324	107	393	23176
10	Random forest [Entropy] - Test	0.993762	0.993762	0.993840	0.993777	15067	19	112	5802
11	Random forest [Entropy] - Train	0.999917	0.999917	0.999917	0.999917	60717	7	0	23276
12	Random forest [Gini] - Test	0.993810	0.993810	0.993889	0.993825	15067	18	112	5803
13	Random forest [Gini] - Train	0.999917	0.999917	0.999917	0.999917	60717	7	0	23276
14	Adaboost - Test	0.993190	0.993190	0.993248	0.993204	15070	34	109	5787
15	Adaboost - Train	0.999917	0.999917	0.999917	0.999917	60717	7	0	23276
16	Gradientboost - Test	0.993429	0.993429	0.993497	0.993443	15068	27	111	5794
17	Gradientboost - Train	0.994500	0.994500	0.994541	0.994509	60359	104	358	23179
18	XGboost - Test	0.993286	0.993286	0.993343	0.993299	15071	33	108	5788
19	XGboost - Train	0.994155	0.994155	0.994181	0.994162	60379	153	338	23130

Best models that have been achieved without SMOTE

1. Random Forest (Entropy)
2. Random Forest (Gini)
3. Decision tree (Hyper-tuned)

Top important Features from best Model:

1. Jobs_Retained
2. Jobs_Created
3. Guaranteed_Approved_Loan
4. ChargedOff_Amount
5. Count_Employees
6. Loan_Approved_Gross
7. Gross_Amount_Disbursed
8. Loan_Term
9. Business
10. Low_Documentation_Loan
11. Demography
12. Revolving_Credit_Line